

DISTRIBUTED CONTENT IDENTIFICATION SYSTEM

Publication number: JP2004500761 (T)

Publication date: 2004-01-08

Inventor(s):

Applicant(s):

Classification:

- **international:** G06F13/00; G06Q10/00; H04L12/58; H04L29/06; G06F13/00; G06Q10/00; H04L12/58; H04L29/06; (IPC1-7): G06F13/00; H04L12/58

- **European:** G06Q10/00F2; H04L12/58F; H04L29/08S14

Application number: JP20010547316T 20001222

Priority number(s): US19990469567 19991222; WO2000US42832 20001222

Also published as:

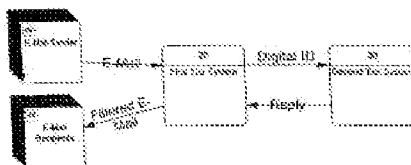
WO0146872 (A1)
US6460050 (B1)
EP1242921 (A1)
EP1242921 (A4)
EP1242921 (B1)
AU4525901 (A)
AT456098 (T)

<< less

Abstract not available for JP 2004500761 (T)

Abstract of corresponding document: **WO 0146872 (A1)**

An e-mail sender (10) transfers an e-mail to its intended recipient (40). The message arrives at a first tier system (20) which may be an e-mail server. A digital identifier engine on the first tier system (20) generates a digital identifier which comprises a hash of at least a portion of the e-mail. Second tier system (30) includes a database and processor which determines, based on an algorithm which varies with the characteristics tested, whether the e-mail is spam or not processes the e-mail accordingly.



.....
Data supplied from the **espacenet** database — Worldwide

(19) 日本国特許庁(JP)

(12) 公表特許公報(A)

(11) 特許出願公表番号

特表2004-500761

(P2004-500761A)

(43) 公表日 平成16年1月8日(2004.1.8)

(51) Int. Cl.⁷

H04L 12/58

G06F 13/00

F I

H04L 12/58

100F

G06F 13/00

610Q

テーマコード (参考)

5K030

審査請求 未請求 予備審査請求 有 (全 30 頁)

(21) 出願番号 特願2001-547316(P2001-547316)
 (86) (22) 出願日 平成12年12月22日(2000.12.22)
 (85) 翻訳文提出日 平成14年6月24日(2002.6.24)
 (86) 国際出願番号 PCT/US2000/042832
 (87) 国際公開番号 W02001/046872
 (87) 国際公開日 平成13年6月28日(2001.6.28)
 (31) 優先権主張番号 09/469,567
 (32) 優先日 平成11年12月22日(1999.12.22)
 (33) 優先権主張国 米国(US)

(71) 出願人 502226874
 ベイス マーク レイモンド
 アメリカ合衆国 カリフォルニア州 94
 402 サン マテオ フィフティーンズ
 アベニュー 42
 (71) 出願人 502226885
 タリー ブルックス キャッシュ
 アメリカ合衆国 カリフォルニア州 94
 402 サン マテオ フィフティーンズ
 アベニュー 42
 (74) 代理人 100059959
 弁理士 中村 稔
 (74) 代理人 100067013
 弁理士 大塚 文昭

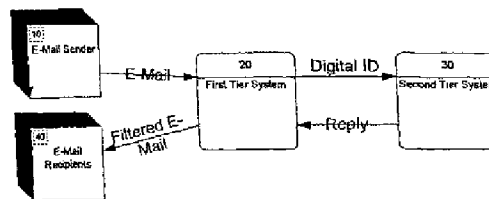
最終頁に続く

(54) 【発明の名称】 分配された内容を識別するシステム

(57) 【要約】

電子メール送信者10は、電子メールを意図した受信者40に対して伝送する。メッセージは、電子メールサーバでありうる第1階層システム20に到達する。第1階層システム20におけるデジタル識別子エンジンは、電子メールの少なくとも一部についてのハッシュを含むデジタル識別子を生成する。第2階層システム30は、データベースと、テストされる特性とともに変化するアルゴリズムに基づいて電子メールがスパムであるか否かを決定しかつこれに従って電子メールを処理するプロセッサと、を含む。

【選択図】 図2



【特許請求の範囲】**【請求項 1】**

デジタル I D 生成部と、
該デジタル I D 生成部から I D を受け取るように接続された I D 出現データベースと、
前記出現データベースにおける I D 状態に基づいた特性を有するとしてファイルを識別する特性比較ルーチンと、
を備えることを特徴とするファイル内容分類システム

【請求項 2】

前記 I D 生成部は、ハッシュアルゴリズムを備える請求項 1 に記載の内容分類システム。

【請求項 3】

前記ハッシュアルゴリズムは、MD5 ハッシュアルゴリズムである請求項 2 に記載の内容分類システム。

【請求項 4】

前記 I D 出現データベースは、デジタル I D の出現の頻度を調べる請求項 1 に記載の内容分類システム。

【請求項 5】

異なるシステムにおける複数のデジタル I D 生成部であって、すべて前記 I D 出現データベースに接続され該 I D 出現データベースに I D を与えるデジタル I D 生成部を備える請求項 1 に記載の内容分類システム。

【請求項 6】

前記複数のデジタル I D 生成部は、私設ネットワークおよび公衆ネットワークの組み合わせを介して、前記データベースに接続されている請求項 5 に記載の内容分類システム。

【請求項 7】

前記データベースは、前記複数の生成部に接続される中間サーバに接続される請求項 6 に記載の内容分類システム。

【請求項 8】

前記中間サーバはウェブサーバである請求項 6 に記載の内容分類システム。

【請求項 9】

前記特性は、ジャンクメールを含み、

前記特性は、デジタル I D の出現の頻度により定められる、請求項 1 に記載の内容分類システム。

【請求項 10】

データファイルの特性を識別する方法であって、

前記データファイルについてのデジタル識別子を生成する工程であって、該デジタル識別子を処理システムに転送する工程と、

転送された前記識別子が他の識別子と一致するかどうかを決定する工程と、

前記決定する工程に基づいて電子メールを処理する工程と、

を備えることを特徴とする方法。

【請求項 11】

前記生成する工程は、前記データファイルの少なくとも一部をハッシュする工程を含む請求項 10 に記載の方法。

【請求項 12】

前記ハッシュする工程は、MD ハッシュを用いる工程を含む請求項 11 に記載の方法。

【請求項 13】

前記生成する工程は、前記データファイルの多数の部分をハッシュする工程を含む請求項 11 に記載の方法。

【請求項 14】

前記データファイルは、電子メールのメッセージであり、前記決定する工程は、該電子メールがスパムであるかどうかを決定する工程を含む請求項 10 に記載の方法。

【請求項 15】

10

20

30

40

50

前記決定する工程は、デジタル I D が生成される単位時間についての割合を調べることにより、前記電子メールをスパムとして識別する請求項 10 に記載の方法。

【請求項 16】

前記生成する工程は、複数のソースシステムにおいて I D を生成する工程を備え、該複数のソースシステムすべては、前記決定する工程を実行する少なくとも 1 つの処理システムに対してネットワークを介して接続されている請求項 10 に記載の方法。

【請求項 17】

前記電子メールを処理する工程は、前記複数のソースシステムに対して、前記決定する工程に基づいて前記電子メールを用いた作用を実行する請求項 16 に記載の方法。

【請求項 18】

電子メールのメッセージにフィルタリングを行う方法であって、
前記メッセージを処理してデジタル識別子を付与する工程と、
前記メッセージが前記特性を有するかどうかを決定するために、該デジタル識別子をデジタル識別子の特性データベースと比較する工程と、
前記比較する工程に基づいて前記メッセージを処理する工程と、
を備えることを特徴とする方法。

【請求項 19】

前記処理する工程は、少なくとも 1 つの第 1 システムで発生し、前記比較する工程は、第 2 システムで発生する請求項 18 に記載の方法。

【請求項 20】

前記処理する工程は、複数の第 1 システムで発生する請求項 19 に記載の方法。

【請求項 21】

前記少なくとも 1 つの第 1 システムおよび第 2 システムは、インターネットにより接続される請求項 19 に記載の方法。

【請求項 22】

前記比較する工程は、特定 I D が時間期間内に発生する頻度を決定する工程と、前記 I D を特性を有するものとして分類する工程と、デジタル識別子を分類された前記 I D と比較する工程と、を備える請求項 18 に記載の方法。

【請求項 23】

分類すべきファイルを有する第 1 システムと、
該第 1 システムにおけるファイル I D 生成部と、
該 I D 生成部により生成される I D を受け取るべく該 I D 生成部に接続された第 2 システムにおけるデータベースと、
該データベースに関して特性を満たすまたは満たさないとして前記 I D を分類する第 2 システムにおける比較ルーチンと、
を備えることを特徴とするファイル内容分類システム。

【請求項 24】

複数の第 1 システムを含み、
該複数の第 1 システムのそれぞれは、前記第 2 システムにおける前記データベースに接続されるそれぞれのファイル I D 生成部を含む請求項 23 に記載のシステム。

【請求項 25】

前記複数の第 1 システムは、インターネットを介して前記第 2 システムに接続される請求項 24 に記載のシステム。

【請求項 26】

前記第 2 システムは、ウェブサーバインタフェースシステムとデータベースシステムとを備え、
該データベースシステムは、前記ウェブサーバシステムによってインターネットから隔離されている請求項 25 に記載のシステム。

【請求項 27】

ネットワークによって接続された第 1 および第 2 コンピュータに対する内容分類システム

10

20

30

40

50

であって、

前記第1コンピュータにおけるクライアントエージェントファイル識別子生成部と、
前記クライアントエージェントから識別子を受信しかつ該クライアントエージェントに
応答を付与する前記第2コンピュータにおける、サーバ比較エージェントおよびデータ構造
と、を備え、

前記クライアントエージェントは、前記サーバ比較エージェントからの応答に基づいて前
記ファイル処理することとを特徴とするシステム。

【請求項28】

インターネットにおけるサービスを提供する方法であって、

前記インターネットにおけるクライアントエージェントを有する複数のシステムからデー
タベースを有するサーバに対するデータを収集する工程と、

前記データベースにおいて収集された情報に関して受信されたデータに特性を付与する工
程と、

内容識別子を前記クライアントエージェントに伝送する工程と、

を備えることを特徴とする方法。

【請求項29】

前記収集する工程は、データファイルについてのデジタル識別子を収集する工程を含む
請求項28に記載の方法。

【請求項30】

前記データファイルは電子メールである請求項28に記載の方法。

【請求項31】

前記特性を付与する工程は、

特定の識別子の収集の頻度を調べる工程と、

前記頻度に基づいて前記データファイルに特性を付与する工程と、

前記特性を記憶する工程と、

収集された識別子を既知の特性付けと比較する工程と、

を含む請求項29に記載の方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、ネットワーク上のファイルのための内容識別 (c o n t e n t i d e n t i f i c a t i o n) の分野に関する。

【0002】

【従来の技術】

インターネットが普及および成長するのに伴って、公衆ネットワークおよび私設ネットワ
ークの両方におけるシステム間で内容を転送することが、急速に増加してきている。

【0003】

【発明が解決しようとする課題】

インターネットによって多くの情報が多数の人々に比較的安価な方法でもたらされてい
る一方、このような普及にはある欠点がある。このような欠点の1つは、特に電子メール
の成長に関連したものであり、一般的には「スパム」電子メールと称される。スパム電子
メールは、受信者にセールスの機会または「すぐに金持ちになる」スキームを利用できる
ようにさせる意図を有した、通常短時間で大量に送信される好ましくない電子メールであ
る。

【0004】

ユーザは、スパムから逃れるために多数の技術に頼ることになる。最も一般的なのは、電
子メールクライアントプログラムに構築されている電子メールフィルタリングを用いて簡
単なフィルタリングを行うことである。このタイプのフィルタリングでは、ユーザは、特
定の言葉、サブジェクトライン (s u b j e c t l i n e) 、出所アドレス、送信者ま
たはその他の変数に基づいてフィルタを設定し、電子メールのクライアントは、送られて

10

20

30

40

50

きた電子メールを受信した際に、または、サーバレベルで、この電子メールを処理し、フィルタが定義されている方式に依存した動作を行う。

【0005】

より精巧な電子メールフィルタリングサービスが、例えば、わずかな手数料でオフサイト（off-site）フィルタリングがリモートサイト（remote site）で実行されるような場所で、確立されてきている。1つのシステムでは、電子メールは、サービスプロバイダにオフサイトで転送されており、自動フィルタリングが、サービスプロバイダによって更新される探索法（heuristic）に基づいてプロバイダ側で発生する。別のシステムでは、オフサイトフィルタリングは、実際に人々が電子メールを通読し電子メールがスパムであるか否かを判断して発生する。他のシステムは、探索法が用いられ、かつ、周期的に人々がサービスに転送された電子メールを実際に検閲しこの電子メールが上述した定義内に「スパム」を含んでいるかを決定するような、ハイブリッドである。これらハイブリッドのサービスでは、人間がランダムな基準に基づいて検閲するので、サービスにより受信される電子メールの全体のうち一部のみを調べるにすぎない。実際に人が電子メールを検閲するようなシステムでは、電子メールがこの電子メールの送信者または受信者に対して守秘義務を有しているかまたは有していない第三者によりチェックされるので、秘密保持問題が生ずる。

10

【0006】

加えて、外部サービス（outside service）に対する添付ファイルを含む電子メール全体を転送することは、高バンド幅（high bandwidth）問題を意味する。なぜならば、このような転送は、特定の電子メールについてのバンド幅を事実上3度増加させるからである。すなわち、1度目は、最初の伝送のときであり、2度目は、サービスへの伝送のときであり、3度目は、最終的な受信者へ再分配するためにサービスからサーバへ戻されるときである。

20

【0007】

さらには、スパムの送信者は、上述したフィルタを避けることにずいぶん精通してきている。ダイナミックアドレッシング（dynamic addressing）方式、非常に長いサブジェクト、および、匿名再ルーティング（anonymous re-routing）サービスが使用されることによって、通常のフィルタリング方式および上述した探索法ベースのサービスであっても、スパム利用者が方法を変化させてきていることに関して、絶えず最新のものを維持していくことが非常に困難になっている。

30

【0008】

インターネットが普及したことについての別の欠点は、インターネットがコンピュータウイルスを非常に多くの人に伝えるための非常に効果的なメカニズムを有していることである。ウイルスの識別は、一般的には、特定の会社における各コンピュータまたはサーバで起動および常駐するプログラム、および、多くの技術を用いて既知のウイルスについてファイルおよび電子メール添付ファイルを定期的にスキャンするプログラムに限定される。

【0009】

よって、本発明の目的は、効率的かつ最新の方式により内容を識別する内容識別システムを提供することである。

40

本発明のさらに別の目的は、分類（classification）システムの他のユーザにより受信される内容に影響を与えて、この内容の特性を決定することである。

【0010】

本発明のさらに別の目的は、受信者の要求によりネットワーク上の所定の伝送の内容の特性を迅速かつ効果的に認識するサービスを提供することである。

本発明のさらに別の目的は、機密方式で上記目的を実現することである。

本発明のさらに別の目的は、低バンド幅で動作するシステムを提供することである。

【0011】

【課題を解決するための手段】

これらの目的およびその他の目的が本発明に与えられる。本発明は、大まかに説明すれば

50

、ファイル内容分類システムを備える。一態様として、このシステムは、デジタルID生成部と、このID生成部からIDを受け取るように接続されたIDデータベースと、を含む。本システムは、さらに、出現データベースにおけるIDの出現に基づいた特性を有するとしてファイルを識別する特性比較ルーチンを含む。

【0012】

特定の実施の形態では、上記ファイルは電子メールであり、本システムは、ハッシュ処理を利用してデジタルIDを生成する。このIDは、ネットワークを介してプロセッサに転送される。このプロセッサは、特性付け工程および決定工程を実行する。この後、このプロセッサは、上記生成部に応答して、特性付け応答に基づく電子メールの処理を可能とする。

10

【0013】

別の態様として、本発明は、データファイルの特性を識別する方法を含む。この方法は、上記データファイルについてデジタル識別子を生成する工程と、このデジタル識別子を処理システムに転送する工程と、転送されたデジタル識別子が他の識別子の特性と一致するかどうかを決定する工程と、前記決定する工程に基づいて電子メールを処理する工程と、を備える。

【0014】

さらに別の態様として、本発明は、インターネットにおけるサービスを提供する方法を備える。この方法は、インターネットにおけるクライアントエージェントを有する複数のシステムからデータベースを有するサーバへのデータを収集する工程と、上記データベースにおいて収集された情報に関して受信されたデータに特性を付ける工程と、内容識別子を上記クライアントエージェントに伝送する工程と、を含む。この態様では、上記特性を付ける工程は、データファイルについてのデジタル識別子を収集する工程を含む。加えて、上記特性を付ける工程は、特定の識別の収集の頻度を調べる工程と、この頻度に基づいて上記データファイルに特性を付ける工程と、特性付けを記憶する工程と、収集された識別子を既知の特性付けと比較する工程と、を含む。

20

【0015】

【発明の実施の形態】

本発明の特定の実施の形態を参照して本発明を説明する。本発明のその他の目的、特徴および効果は、本明細書および図面を参照することにより明らかであろう。

30

【0016】

本発明は、分類しようとする各内容についてデジタル識別子 (i d e n t i f i e r) を用いかつこのIDに基づいて該内容を特徴付ける、分散内容分類システム (d i s t r i b u t e d c o n t e n t c l a s s i f i c a t i o n s y s t e m) を提供する。本システムの1つの態様では、デジタル識別子は、内容についての特定の特性が存在するかどうかを決定するために、処理アルゴリズムによってその他の任意数の識別子と相関がある処理システムに転送される。本質的には、分類は、この分類が求められる問合せ (q u e r y) に基づいた内容について正誤テストである。例えば、システムは、電子メールがスパムであるかどうかを識別する、または、特定のファイルにおける内容が、著作権で保護されたものであるか否かもしくはウィルスを含んでいるか否かを示す所定の判定基準に一致するかどうかを識別することができる。

40

【0017】

電子メールのメッセージを分類することに関して本発明を説明するが、当業者であれば、本発明のデータ分類システムを、システムに存在するかまたはシステムを通して伝送される任意の種類のテキストもしくはバイナリデータを分類するためにも利用できる、ということを理解できよう。

【0018】

図1は、電子メールの送信者10が、該送信者に転送される前にフィルタリング処理システム15により傍受される (i n t e r c e p t e d) 電子メールを送信するような、本発明を高水準で表現した図である。このシステムは、受信者20がメッセージを見る前に

50

電子メールに対して作用する能力を有する。

【0019】

図2は、電子メール送信者10が意図した受信者40に対して電子メールを送信し、このメッセージがこの例では電子メールサーバを示す第1階層(tier)システム20に到達する際における、電子メールの内容における本発明の概略的な処理を示す。通常(本発明のシステムが存在しないとき)、第1階層システム20が意図された受信者に直接電子メールを送信するのは、この受信者の電子メールクライアントアプリケーションがこの電子メールの伝送を要求した際である。本発明では、電子メールサーバと協同する第1階層システムにおけるデジタル識別子エンジンが、1つの環境では上記電子メールの少なくとも一部のハッシュ(hash)を備えるデジタル識別子を発生させる。この後、デジタル識別子は、第2階層システム30に転送される。第2階層システム30は、データベースと、テストされる特性とともに変化するアルゴリズムに基づいて、電子メールが上記問合せ(例えば電子メールがスパムであるか否か?)の分類を満たすかどうかを決定するプロセッサと、を含む。

10

【0020】

このアルゴリズムの結果に基づいて、第2階層システム30から第1階層システム20に対して応答が送信される。ここで、第1階層システム20は、フィルタの結果に基づいてユーザにより再生成された記述(description)に従って電子メールを処理する。この結果については図2に示されていることからわかるように、フィルタリングされた電子メールプロダクトは、電子メールの受信者に転送される。以下、第2階層システム30で計算されたアルゴリズムの結果に依存する電子メールの処理についてのその他のオプションを説明する。

20

【0021】

図1および図2を参照すれば、外部の電子メール送信者については、システム外部のソースからフィルタリング処理に送られた電子メールまたは電子データの任意のソースとすることができる。電子メール受信者40は、フィルタリング処理を通過する電子データの最終的な目的地を示す。

【0022】

一態様では、第1階層システム20上で動作しかつhttp://www.w3.org/TR/1998/Rec-Dsig-label/MD5_1_6で完全に記載されたMD5ハッシュに従ってデジタルIDを発生させる実行可能コードにより、システムを実行することができる。しかしながら、任意のハッシュアルゴリズムを利用できることを認識されたい。一実施の形態では、MD5ハッシュにより生成されたデジタルIDは、サブジェクトライン全体のうち2つのスペースが現れる部分までであり、本体の全体およびメッセージの本体における最後の500バイトである。さらに、生成されるデジタルIDについては1つのハッシュまたは多数のハッシュとすることができ、ハッシュアルゴリズムについては考慮するデータの全部分またはいくつかの部分に対して実行することができる、ということを理解されたい。例えば、サブジェクトラインのうち、該サブジェクトラインにおけるいくつかの数の文字、メッセージ本体のすべてまたは一部を、ハッシュとすることができる。さらに、デジタルIDは固定長とする必要がないことを認識されたい。

30

40

【0023】

実行可能な第1階層は、分離処理(separate process)として、または、第1階層システム20上で動作する電子メールシステムについてのプラグイン(plugin)として、動作することができる。一実施の形態では、第1階層システム20のような動作システム上で通常用いられるメールサーバを有する実行可能なインタフェースは、SendmailTMとして知られている。SendmailTMとともに利用される一般的なセットのツールは、Procmail(http://www.iinet.com/internet/robots/procmail)である。本発明のシステムの一態様では、実行可能なものは、SendmailTMおよびProcmailである。こ

50

のような実施の形態では、コンフィギュレーションファイル（sendmail.cfのような）は、デジタルIDを生成して第2階層システムに伝送し、その応答を受信し、この応答メッセージの結果としてメッセージに対する処理または削除をProcmailに指示するのに実行可能な第1階層サイト電子メールを介して、受信されてくる電子メールを処理するように、Procmailサーバプログラムに対して指示する、1行のコードを含む。

【0024】

この実行可能なもの（executable）については、例えばパール（perl）スクリプトにより記述することができ、商用のシステムもしくは無料電子メールシステム、または、電子メール以外のアプリケーションにおけるその他のデータ転送システムと相互作用するようにデザインすることができる、ということを理解されたい。

10

【0025】

このコンテキストにおけるデジタルIDを利用することにより、ネットワークをわたって第2階層システムに伝送するのに必要なバンド幅を低減することができる。このIDは、典型的には、ハッシュデータだけでなく第1階層システム20上で動作する実行可能なもののタイプを第2階層システム30に通知するバージョン情報をも含む。

【0026】

加えて、第2階層システムの第1階層システムに対する応答については、例えば、第1階層システムがかかる要求をする権限を有していない場合に第2階層システム30から第1階層システム20へのサービスの拒絶とすることができる。本発明によれば、ボリュームまたはその他の収益判定基準に基づく手数料に対して、フィルタリングサービス（すなわち、第2階層サービス処理を運営し、第2階層データベースを維持すること）を提供することにより、収益をあげることができる、ということを認識されたい。この商用コンテキストでは、応答については、所定の期間に対して割り当てられたフィルタリングクォータ（quota）を上回った、第1階層システムのユーザについてのサービスの拒絶とすることができる。

20

【0027】

図3は、本発明のシステムの第2の実施の形態を示す。図3では、第1階層システムは、メッセージ予備処理部110、メッセージ処理部120およびコンフィギュレーションファイルDS10を含む3つの構成要素に分割される。この例では、送信者10からの電子メールは、最初にメッセージ予備処理部110に迂回する。予備処理アルゴリズムは、コンフィギュレーションファイルDS10からの規則を用いて構成される。これらの規則は、例えば受信される電子メールからデジタルIDをいつどのように生成するかについてのガイドラインである。メッセージ処理部は、電子メールの送信者10から電子メールを受信し、DS10からの処理規則に基づいてデジタルIDを生成する。DS10は、第1階層システム20についてのコンフィギュレーション規則（予備処理および後処理の前の）を記憶するコンフィギュレーションファイルである。メッセージ処理規則は、スパムとして分類された電子メールをどのように処理するかについてのガイドラインを含むことができる。例えば、メッセージが検出され、このメッセージは、第2階層システム30によりスパムであるとみなされかつサブジェクトラインに付加された「SPAM」という言葉（word）を有する電子メールのための保持領域に転送され、分離フォルダおよびこれと同等のものに移動しうる。この例では、メッセージ予備処理規則は、システムのフィルタリングにより、特定の目的地すなわちアドレスからのすべての電子メールを除去しうる規則を含む。メッセージがこのような除去についての判定基準を満たせば、メッセージは、ライン50に示すように、電子メールの受信者40に直接転送するためのメッセージ処理部120に、自動的に直接転送される。このような規則は、電子メールを拒絶メッセージ保管部DS20に直接転送するための判定基準を含む。

30

40

【0028】

予備処理規則が、特定の電子メールがシステム中を直接通過することを示さない場合には、1つ以上のデジタル識別子がライン66に示すように生成されて第2階層システム3

50

0 に伝送される。図 3 に示すような例では、第 2 階層システム 30 は、第 3 階層データベース 220 における第 2 階層サーバ 210 を含む。この例では、第 2 階層サーバは、デジタル ID を中継し、予備処理部とメッセージ処理部 120 との間で応答する。図 3 に示す例は、第 2 階層サーバ 210 がインターネットを介してアクセス可能なウェブサーバを備えかつ第 3 階層データベース 220 が一連のファイアウォールまたはその他のセキュリティ手段を用いて第 2 階層サーバによりインターネットから遮断されているような、インターネットベースの環境で特に有用である。これにより、第 3 階層データベースで編集されるデジタル ID 情報のデータベースは、このシステムのセキュリティを危険にさらすことを望む者による攻撃から確実に免れることができる。

【0029】

この場合、第 2 階層サーバ 210 は、当該データをテストするためのアルゴリズムに基づいて ID を処理する第 3 階層データベースに対して直接デジタル ID を転送する。第 3 階層データベースは、第 2 階層サーバによりメッセージ処理部 120 に戻される応答を生成する。この後、メッセージ処理部 120 は、フィルタリングされた電子メールを電子メールの受信者に送ること、フィルタリングされた電子メールを拒絶メッセージ保管部 DS20 に送信すること、または、コンフィギュレーションファイル DS10 に特定されたユーザ選択コンフィギュレーション設定に従ってメッセージに作用すること、のいずれかにより、電子メールに作用する。

【0030】

図 3 に示す環境では、第 1 階層におけるコンフィギュレーションファイル DS110 によって、電子メールの送信者から受信した電子メールについてのその他の判断を、第 2 階層 30 からの応答に基づいて行うことができる。例えば、スパム電子メールを削除するだけでなく、電子メールが「スパム」であるということを示すようにサブジェクトラインを追加することができ、いくらかの時間だけ、生成された自動応答およびこれと同等のものについての隔離ゾーンにこの電子メールを保持しておくことができる。加えて、メッセージ予備処理およびメッセージ処理規則によって、第 2 階層システム 30 がアクセス不可能であるような状況を考慮するように電子メール処理に対する決定を行うことができる。このような場合に実行できる決定には、「すべての電子メールを転送する」、「電子メールを転送しない」、「さらなる処理を保留する」およびこれらと同等のものが含まれる。

【0031】

インターネットベースの環境では、第 2 階層サーバ 30 は、HTTP プロトコルの手段により、第 3 階層データベース 220 に対して、デジタル識別およびその他の情報を伝送することができる。本発明によればその他のプロトコルを利用できるということを認識されたい。第 3 階層データベース 220 は、任意数の様々な商用データベースプラットフォームにおいて維持することができるものである。加えて、第 3 階層データベースは、クライアント識別子トラッキング (tracking) のようなシステム管理情報と、収益処理情報とを含むことができる。本発明の固有の態様では、第 3 階層データベース 22 におけるデジタル ID は、汎用的なベースで維持される。すなわち、第 2 階層サーバ 210 に対してデジタル ID を送信するすべての第 1 階層サーバは、データベースに対するデータと、第 3 階層システム上で動作する処理アルゴリズムとに寄与する。スパムの決定が目標であるような一実施の形態では、アルゴリズムが、例えばメッセージ（または実際にはメッセージについての ID）が特定のタイムフレーム内で受信される頻度を計算することができる。例えば、同一のメッセージを示す特定の ID が単位時間あたりに数回見受けられる場合には、システムは、このメッセージ（および ID）をスパムとして分類する。スパムとして分類されたこの ID と一致する今後のすべての ID によって、ここでは、システム 30' は、電子メールがスパムであるという応答を生成することになる。本発明のシステムに加わる第 1 階層システム 20' を有する各クライアントは、その他のクライアントにより生成されるデータから利益を受ける。よって、例えば、特定のクライアントは、この後同様のメッセージを見る第 1 階層システム 20' を有する別のクライアントをシステムに対して分類させてしまうような頻度の条件を満たす多数の電子メールを受信した

10

20

30

40

50

場合には、メッセージがスパムであるという応答を自動的に受信するであろう。

【0032】

特定の場合には、評判の良い大企業は、例えば電子メール受信者により具体的に要求される情報メーリングリストサーバのような広範囲な数のユーザに対して、大きなブロックの電子メールを転送するということを認識されたい。本システムは、エリアおよび第2階層システムレベルの両方におけるこのようなメーリングリストアプリケーションを考慮している。評判の良いサーバは、目的地であるシステム20'において多数の受信者に対して多数の電子メールを送信できなくてはならないという事実を考慮するために、第3階層データベース220上で動作するアルゴリズムに例外を設けることができる。これに代えて、または、このような例外とともに、ユーザは、DS10コンフィギュレーションを介して自分専用の例外を定義することができる。サービスとして、例えばフォーチュン1000カンパニーのドメインネームのような任意数の受け入れられるソースを、免除された「スパムでない」サイトとして特長付けることができ、ユーザは、サーバ側の設定を「信用する」または「信用しない」ように選択することができる。

10

【0033】

上述した実施の形態は、メッセージがスパムであるかどうかを決定するために頻度のアルゴリズムを利用しているが、このアルゴリズムにおけるさらなる実施の形態では、特定の文字または言葉の頻度についてメッセージを分析し、およびまたは、特定のメッセージにおける最も一般的な言葉の2番目に一般的な言葉に対する関係を分析することができる。このアルゴリズムの任意数の変形を用いることができる。

20

【0034】

さらに、ユーザをさらなるメーリングリストおよび参照ソースに接続すること、受信者の特性をその他の者にフィードバックすること、および、これらと同等ことのような、付加価値サービスに調和するように、第2階層サーバを利用することができる、ということを認識されたい。

【0035】

図4は、本発明の別の実施の形態およびサーバ側システムがデジタル識別子をどのように処理するのかを示す。図4において、本実施の形態は、メッセージ免除判定基準およびメッセージ処理規則をそれぞれメッセージ予備処理部110およびメッセージ処理部120に付与するDS10コンフィギュレーションファイルを含む。メッセージ処理部110については、2つの構成要素、すなわち、メッセージ免除チェック111およびデジタルID生成112として考えることができる。これらの構成要素の両方は、図3を参照して説明したように、免除電子メールを電子メールの受信者40に直接伝えることを可能とするように、または、デジタルIDを第2階層サーバ210に転送する必要があるかどうかを決定するように機能する。メッセージ処理アルゴリズム120により受信される応答は、規則決定アルゴリズム121および電子メールフィルタリング123により作用を受ける。

30

【0036】

第2階層システム30'では、第2階層プロセッサ210から伝送されたデジタルIDは、デジタルIDプロセッサ221に伝送される。この実施の形態では、プロセッサ221は、単位時間ごとに各デジタルIDについてDS30に記憶されるカウンタデータを増加させる。データベース220により処理されるメッセージのボリュームは非常に大きくなりうるので、システムの全メッセージボリュームの百分率としてみられる各メッセージのボリュームの変化を認識するように、頻度アルゴリズムを調整することができる。

40

【0037】

DS30に格納される頻度データは、生成され第2階層サーバ210に転送された応答がメッセージがスパムであるか否かを示すべきであるかどうかを、DS30におけるデータと所定のクライアントについての特定の情報とに基づいて決定する応答生成部222に、送られる。コンフィギュレーションファイルDS40は、上述したように、第2階層サーバ210からの応答がメッセージプロセッサ120における規則決定部（この規則決定部

50

は、上述したように、スパムであるということが実際に決定された場合にどのように規則を処理するかを決定する)に転送されるということを示す規則を含むことができる。フィルタリングされた電子メール分配アルゴリズムは、電子メールを電子メールクライアント40または上述した拒絶メッセージ保管部に直接転送する。

【0038】

本発明の重要な特徴は、データ識別子保管部DS30において用いられるデジタルIDは、多数の異なる第1階層システムから引き出されるということである。よって、第2階層サーバおよび後段のデータベース220に接続される第1階層システムの数が多くなるほど、システムはより効果的となる。

【0039】

その他のアプリケーションが、スパム電子メールを検出することだけでなく、ネットワークを介して伝送されたウィルスを検出すること、および、同様に伝送された著作権により保護されたものを識別することを含むことができる、ということを示すことができる。さらには、処理デジタル識別子およびデータ記憶DS30についてのアルゴリズムは静的なものではなく、このアルゴリズムについては、頻度以外に、テストされるメッセージまたはデータのその他の特性を見つけるように調整することができる、ということを示すことができる。

【0040】

よって、本システムは、機密かつ効果的な電子メールフィルタを提供する一方で限定された量のバンド幅を利用するフィルタシステムを提供するようにデータベースに接続された、第1階層システムまたはクライアントの数の間に影響を与える(leveraging)ことができる。さらに、第2および第3階層システムを維持することによって、第2階層システムの処理を提供するサービスについての手数料を課金することにより得られるサービスに対する利益を発生させることができるということを示すことができる。

【0041】

さらに、本システムは、分類される内容に関する匿名統計データ(anonymous statistical data)を収集および頒布することができる。例えば、電子メールフィルタリングがシステムの主なアプリケーションであるところでは、システムは、フィルタリングされる全電子メールのうちスパムが占める百分率、このようなスパムが発生する場所、および、これらと同等のことを識別し、識別したものを利害関係者(interested parties)に対して手数料またはその他の報酬(compensation)と引き換えに頒布することができる。

【図面の簡単な説明】

【図1】

従来技術にかかる内容を識別すべく電子メールに対してフィルタリングする際のシステムを示すブロック図

【図2】

本発明の処理を示すブロック図

【図3】

本発明の方法および装置を詳細に示すブロック図

【図4】

本発明の方法および処理の第2の実施の形態を示すブロック図

10

20

30

40

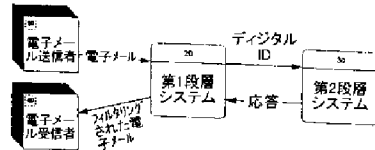
【 図 1 】

Figure 1



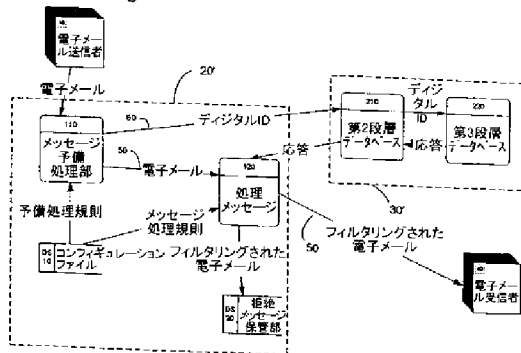
【图 2】

Figure 2



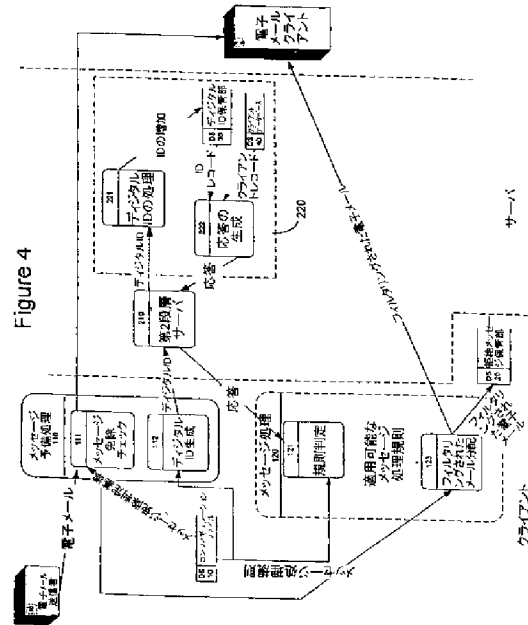
【 図 3 】

Figure 3



【 図 4 】

Figure 4



【国際公開パンフレット】

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

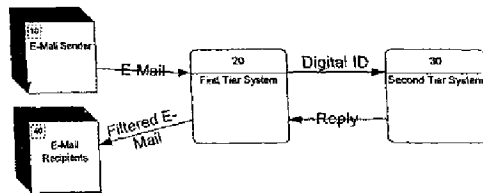
(19) World Intellectual Property Organization
International Bureau(43) International Publication Date
28 June 2001 (28.06.2001)

PCT

(10) International Publication Number
WO 01/46872 A1

- (51) International Patent Classification: G06F 17/30 (81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MY, NZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.
- (21) International Application Number: PCT/US01/42032 (84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).
- (22) International Filing Date: 22 December 2000 (22.12.2000)
- (25) Filing Language: English
- (26) Publication Language: English
- (50) Priority Data: 09/469,567 22 December 1999 (22.12.1999) US
- (71) Applicants and (72) Inventors: PACE, Mark, Raymond (US/US); 42 15th Avenue, San Mateo, CA 94402 (US); TALLEY, Brooks, Cash (US/US); 42 15th Avenue, San Mateo, CA 94402 (US).
- (74) Agents: VIDERA, Larry, E. et al.; Flisler Dubb Meyer and Lovejoy LLP, Four Embarcadero Center, Suite 409, San Francisco, CA 94111-4156 (US).
- Published:
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments.
For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: DISTRIBUTION CONTENT IDENTIFICATION SYSTEM



(57) Abstract: An e-mail sender (10) transfers an e-mail to its intended recipient (40). The message arrives at a first tier system (20) which may be an e-mail server. A digital identifier engine on the first tier system (20) generates a digital identifier which comprises a hash of at least a portion of the e-mail. Second tier system (30) includes a database and processor which determines, based on an algorithm which varies with the characteristics tested, whether the e-mail is spam or not processes the e-mail accordingly.

WO 01/46872 A1

WO 01/46872

PCT/US00/42832

DISTRIBUTED CONTENT IDENTIFICATION SYSTEM

BACKGROUND OF THE INVENTIONField of the Invention

The invention relates to the field of content identification for files on a network.

5

Description of the Related Art

With the proliferation and growth of the Internet, content transfer between systems on both public and private networks has increased exponentially. While the Internet has brought a good deal of information to a large number of people in a relatively inexpensive manner, this proliferation has certain downsides. One such downside, associated with the growth of e-mail in particular, is generally referred to as "spam" e-mail. Spam e-mail is unsolicited e-mail which is usually sent out in large volumes over a short period of time with the intent of inducing the recipient into availing themselves of sales opportunities or "get rich quick" schemes.

15

To rid themselves of spam, users may resort to a number of techniques. The most common is simple filtering using e-mail filtering which is built into e-mail client programs. In this type of filtering, the user will set up filters based on specific words, subject lines, source addresses, senders or other variables, and the e-mail client will process the incoming e-mail when it is received, or at the server level, and take some action depending upon the manner in which the filter is defined.

20

More elaborate e-mail filtering services have been established where, for a nominal fee, off-site filtering will be performed at a remote site. In one system, e-mails are forwarded offsite to a service provider and the automatic filtering occurs at the provider's location based on heuristics which are updated by the service provider. In other systems, offsite filtering occurs using actual people to read through e-mails and judge whether e-mail is spam or not. Other systems are hybrids, where heuristics are used and, periodically, real people review e-mails which are forwarded to the service to determine whether the e-mail constitutes "spam" within the aforementioned definition. In these hybrid services, personal reviews occur on a random basis and hence constitute only a spot check of the entire volume of e-mail which is received by the service. In systems where real people review e-mails, confidentiality issues arise since e-mails are reviewed by a third party who may or may not be under an obligation of confidentiality to the sender or recipient of the e-

25

30

WO 01/46872

PCT/US00/42832

mail.

In addition, forwarding the entire e-mail including attachments to an outside service represents a high bandwidth issue since effectively this increases the bandwidth for a particular e-mail by three times: once for the initial transmission, the second time for the transmission to the service and the third time from the service back to server for redistribution to the ultimate recipient.

Further, senders of spam have become much more sophisticated at avoiding the aforementioned filters. The use of dynamic addressing schemes, very long-length subject lines and anonymous re-routing services makes it increasingly difficult for normal filtering schemes, and even the heuristics-based services discussed above, to remain constantly up-to-date with respect to the spammers' ever changing methods.

Another downside to the proliferation of the Internet is that it is a very efficient mechanism for delivering computer viruses to a great number of people. Virus identification is generally limited to programs which run and reside on the individual computer or server in a particular enterprise and which regularly scan files and e-mail attachments for known viruses using a number of techniques.

SUMMARY OF THE INVENTION

Hence, the object of the invention is to provide a content classification system which identifies content in an efficient, up-to-date manner.

The further object of the invention is to leverage the content received by other users of the classification system to determine the characteristic of the content.

Another object of the invention is to provide a service which quickly and efficiently identifies a characteristic of the content of a given transmission on a network at the request of the recipient.

Another object of the invention is to provide the above objects in a confidential manner.

A still further object is to provide a system which operates with low bandwidth.

These and other objects of the invention are provided in the present invention. The invention, roughly described, comprises a file content classification system. In one aspect the system includes a digital ID generator and an ID database coupled to receive IDs from the ID generator. The system further includes a characteristic comparison routine identifying the file as having a characteristic based

WO 01/46872

PCT/US90/42832

on ID appearance in the appearance database.

In a particular embodiment, the file is an e-mail file and the system utilizes a hashing process to produce digital IDs. The IDs are forwarded to a processor via a network. The processor performs the characterization and determination steps. The processor then replies to the generator to enable further processing of the email based on the characterization reply.

In a further aspect, the invention comprises a method for identifying a characteristic of a data file. The method comprises the steps of: generating a digital identifier for the data file and forwarding the identifier to a processing system; determining whether the forwarded identifier matches a characteristic of other identifiers; and processing the e-mail based on said step of determination.

In yet another aspect, the invention comprises a method for providing a service on the Internet, comprising: collecting data from a plurality of systems having a client agent on the Internet to a server having a database; characterizing the data received relative to information collected in the database; and transmitting a content identifier to the client agent. In this aspect, said step of collecting comprises collecting a digital identifier for a data file. In addition, said step of characterizing comprises: tracking the frequency of the collection of a particular identifier; characterizing the data file based on said frequency; storing the characterization; and comparing collected identifiers to the known characterization

BRIEF DESCRIPTION OF THE DRAWINGS

The invention will be described with respect to the particular embodiments thereof. Other objects, features, and advantages of the invention will become apparent with reference to the specification and drawings in which:

Figure 1 is a block diagram indicating the system in filtering e-mail to identify content in accordance with the prior art.

Figure 2 is a block diagram illustrating the process of the present invention.

Figure 3 is a block diagram illustrating in additional detail the method and apparatus of the present invention.

Figure 4 is a block diagram illustrating a second embodiment of the method and apparatus of the present invention.

DETAILED DESCRIPTION

The present invention provides a distributed content classification system which

WO 01/46872

PCT/US00/42832

utilizes a digital identifier for each piece of content which is sought to be classified, and characterizes the content based on this ID. In one aspect of the system, the digital identifier is forwarded to a processing system which correlates any number of other identifiers through a processing algorithm to determine whether a particular characteristic for the content exists. In essence, the classification is a true/false test for the content based on the query for which the classification is sought. For example, a system can identify whether a piece of e-mail is or is not spam, or whether the content in a particular file matches a given criteria indicating it is or is not copyrighted material or contains or does not contain a virus.

While the present invention will be discussed with respect to classifying e-mail messages, it will be understood by those of average skill in the art that the data classification system of the present invention can be utilized to classify any sort of text or binary data which resides on or is transmitted through a system.

Figure 1 is a high level depiction of the present invention wherein an e-mail sender 10 transmits an e-mail which is intercepted by a filtering process/system 15 before being forwarded to the sender. The system has the ability to act on the e-mail before the recipient 20 ever sees the message.

Figure 2 illustrates the general process of the present invention in the e-mail context when an e-mail sender 10 transfers an e-mail to its intended recipient 40, the message arrives at a first tier system 20 which in this example may represent an e-mail server. Normally (in the absence of the system of the present invention), the first tier system 20 will transmit an e-mail directly to the intended recipient when the recipient's e-mail client application requests transmission of the e-mail. In the present invention, a digital identifier engine on the first tier system cooperating with the e-mail server will generate a digital identifier which comprises, in one environment, a hash of at least a portion of the e-mail. The digital identifier is then forwarded to a second tier system 30. Second tier system 30 includes a database and processor which determines, based on an algorithm which varies with the characteristic tested, whether the e-mail meets the classification of the query (e.g. is it spam or not?).

Based on the outcome of this algorithm, a reply is sent from the second tier system 30 to the first tier system 20, where the system then processes the e-mail in accordance with the regenerated description by the user based on the outcome of the filter. The result can be as shown in Figure 2, the filtered e-mail product being forwarded to the e-mail recipient. Other options for disposition of the e-mail depending upon the outcome of the algorithm computed at second tier system 30 are described below.

WO 01/46872

PCT/US00/42832

It should be understood with reference to Figures 1 and 2 that the external e-mail sender can be any source of electronic mail or electronic data sent to the filtering process from sources outside the system. The e-mail recipients 40 represent the final destination of electronic data that passes through the filtering process.

5 In one aspect, the system may be implemented in executable code which runs on first tier system 20 and generates digital IDs in accordance with the MD5 hash fully described at http://www.w3.org/TR/1998/Rec-DSig-label-MD5_1_6. It should be recognized however that any hashing algorithm can be utilized. In one embodiment, the digital ID generated by the MD5 hash is of the entire subject line up to the point where two
10 spaces appear, the entire body, and the last 500 bytes of the body of the message. It should be further understood that the digital ID generated may be one hash, or multiple hashes, and the hashing algorithm may be performed on all or some portion of the data under consideration. For example, the hash may be of the subject line, some number of characters of the subject line, all of the body or portions of the body of the message. It
15 should further be recognized that the digital ID is not required to be of fixed length.

The first tier executable may be run as a separate process or as a plug-in with the e-mail system running on a first tier system 20. In one embodiment, the executable interfaces with a commonly used mail server on a running system such as a first tier system 20 is known as Sendmail™. A common set of tools utilized with Sendmail™ is Procmail. (<http://www.ll.com/internet/robots/procmail>). In one aspect of the system of the present invention, the executable may interface with Sendmail™ and Procmail. In such an
20 embodiment, a configuration file (such as a sendmail.cf) includes a line of code which instructs the Procmail server program to process incoming e-mails through the first tier site e-mail executable to generate and transport digital IDs to the second tier system, receive its reply, and instruct the Procmail to process or delete the message, as a result of the reply
25 message.

It should be understood that the executable may be written in, for example, perl script and can be designed to interact with any number of commercial or free e-mail systems, or other data transfer systems in applications other than e-mail.

30 The digital ID usage in this context reduces bandwidth which is required to be transported across the network to the second tier system. Typically, the ID will not only contain the hashed data, but may include versioning information which informs the second tier system 30 of the type of executable running on the first tier system 20.

In addition, the reply of the second tier system to the first tier system may be, for
35 example, a refusal of service from the second tier system 30 to the first tier system 20 in

WO 01/46872

PCT/US00/42832

cases where the first tier system is not authorized to make such requests. It will be recognized that revenue may be generated in accordance with the present invention by providing the filtering service (i.e. running the second tier service process and maintaining the second tier database) for a fee based on volume or other revenue criteria. In this commercial context, the reply may be a refusal of service of the user of the first tier system 20 which has exceeded their allotted filtering quota for a given period.

Figure 3 shows a second embodiment of the system of the present invention. In Figure 3, the first tier system is broken down into three components including a message preprocessing section 110, a message processing section 120, a configuration file DS10. In this example, the e-mail from sender 10 is first diverted to message preprocessing 110. Preprocessing algorithm is configured with rules from configuration file DS10. These rules are guidelines on how and when, for example, to generate digital IDs from the e-mail which is received. Message preprocessing receives the email from the e-mail sender 10 and generates digital IDs based on the preprocessing rules from DS10. DS10 is a configuration file which stores configuration rules (before preprocessing and postprocessing) for the first tier system 20. The message processing rules may include guidelines on how to dispose of those e-mails classified as spam. For example, a message may be detected, and may be forwarded to a holding area for electronic mail that has been deemed to be spam by second tier system 30, have the word "SPAM" added to the subject line, moved to a separate folder, and the like. In this example, message preprocessing rules include rules which might exempt all e-mails from a particular destination or address from filtering by the system. If a message meets such exemption criteria, the message is automatically forwarded, as shown on line 50, directly to message processing 120 for forwarding directly onto the e-mail recipient 40. Such rules may also comprise criteria for forwarding an e-mail directly to a rejected message depository DS20.

If a preprocessing rule does not indicate a direct passage of a particular e-mail through the system, one or more digital identifiers will be generated as shown at line 60 and transmitted to the second tier system 30. In the example shown at Figure 3, second tier system 30 includes a second tier server 210 in a third tier database 220. In this example, the second tier server relays digital IDs and replies between preprocessing and the message processing 120. The example shown in Figure 3 is particularly useful in an Internet based environment where the second tier server 210 may comprise a web server which is accessible through the Internet and the third tier database 220 is shielded from the Internet by the second tier server through a series of firewalls or other security measures. This ensures that the database of digital ID information which is compiled at the third tier

WO 01/46872

PCT/US00/42832

database 220 is free from attack from individuals desirous of compromising the security of this system.

In this case, second tier server 210 forwards the digital ID directly to the third tier database which processes the IDs based on the algorithm for testing the data in question. The third tier database generates a reply which is forwarded by the second tier server back to message processing 120. Message processor 120 can then act on the e-mail by either sending filtered e-mail to the e-mail recipient, sending the filtered e-mail to the rejected message depository DS20 or acting on the message in accordance with user-chosen configuration settings specified in configuration file DS10.

In the environment shown in Figure 3, the configuration file DS110 on the first tier allows other decisions about the e-mail received from the e-mail sender 10 to be made, based on the reply from second tier 30. For example, in addition to deleting spam e-mail, the subject line may be appended to indicate that the e-mail is "spam," the e-mail may be held in a quarantine zone for some period of time, an auto reply generated, and the like. In addition, the message preprocessing and message processing rules allow decisions on e-mail processing to account for situations where second tier system 30 is inaccessible. Decisions which may be implemented in such cases may include "forward all e-mails," "forward no e-mails," "hold for further processing," and the like.

In an internet based environment, the second tier server 30 may transmit a digital identification and other information to the third tier database 220 by means of the HTTP protocol. It should be recognized that other protocols may be used in accordance with the present invention. The third tier database 220 may be maintained on any number of different commercial database platforms. In addition the third tier database may include system management information, such as client identifier tracking, and revenue processing information. In an unique aspect of the present invention in general, the digital IDs in third tier database 22 are maintained on a global basis. That is, all first tier servers which send digital IDs to second tier servers 210 contribute data to the database and the processing algorithm running on the third tier system. In one embodiment, where spam determination is the goal, the algorithm computes, for example, the frequency with which a message (or, in actuality, the ID for the message), is received within a particular time frame. For example, if a particular ID indicating the same message is seen some number of times per hour, the system classifies the message (and ID) as spam. All subsequent IDs matching the ID classified as spam will now cause the system 30 to generate a reply that the e-mail is spam. Each client having a first tier system 20' which participates in the system of the present invention benefits from the data generated by other clients. Thus, for example, if

WO 01/46872

PCT/US00/42832

a particular client receives a number of spam e-mails meeting the frequency requirement causing the system to classify another client having a first tier system 20' which then sees a similar message will automatically receive a reply that the message is spam.

It should be recognized that in certain cases, large reputable companies forward a large block of e-mails to a widespread number of users, such as, for example information mailing list servers specifically requested by e-mail receivers. The system accounts for such mailing list application on both the area and second tier system levels. Exceptions may be made in the algorithm running on the third tier database 220 to take into account the fact that reputable servers should be allowed to send a large number of e-mails to a large number of recipients at the destination system 20'. Alternatively, or in conjunction with such exceptions, users may define their own exceptions via the DS10 configuration. As a service, any number of acceptable sources such as, for example, the Fortune 1000 companies' domain names may be characterized as exempted "no spam" sites, and users can choose to "trust" or "not trust" server side settings.

While the aforementioned embodiment utilizes a frequency algorithm to determine whether a message is spam, additional embodiments in the algorithm can analyze messages for the frequency of particular letters or words, and/or the relationship of the most common words to the second most common words in a particular message. Any number of variants of the algorithm may be used.

It should be further recognized that the second tier server can be utilized to interface with the value added services, such as connecting the users to additional mailing lists and reference sources, providing feedback on the recipients' characteristics to others, and the like.

Figure 4 shows a further embodiment of the invention and details how the server side system manipulates with the digital identifiers. In Figure 4, the embodiment includes a DS10 configuration file which provides message exemption criteria and message processing rules to both message preprocessing and routine 110 message processing 120, respectively. Message preprocessing 110 may be considered as two components: message exemption checking 111 and digital ID creation 112. Both of these components function as described above with respect to Figure 3 allowing for exempt e-mails to be passed directly to an e-mail recipient 40, or determining whether digital IDs need to be forwarded to second tier server 210. Replies are received by message processing algorithm 120 is acted on by rule determination algorithm 121, and e-mail filtering 123.

At the second tier system 30', digital IDs transmitted from second tier processor 210 are transmitted to a digital ID processor 221. In this embodiment, processor 221

WO 01/46872

PCT/US00/42832

increments counter data stored in DS30 for each digital ID per unit time. As the volume of messages processed by database 220 can be quite large, the frequency algorithm may be adjusted to recognize changes in the volume of individual messages seen as a percentage of the total message volume of the system.

5 The frequency data stored at DS30 feeds a reply generator 222 which determines, based on both the data in the DS30 and particular information for a given client, (shown as data record DS40) whether the reply generated and forwarded to second tier server 210 should indicate that the message is spam or not. Configuration file DS40 may include rules, as set forth above, indicating that the reply from the second tier server 210 is
10 forwarded to rule determination component of message processor 120 which decides, as set forth above, how to process the rule if it is in fact determined that it is spam. The filtered e-mail distribution algorithm forwards the e-mail directly to the e-mail client 40 or to the rejected message repository as set forth above.

15 A key feature of the present invention is that the digital IDs utilized in the data identifier repository DS30 are drawn from a number of different first tier systems. Thus, the greater number of first tier systems which are coupled to the second tier server and subsequent database 220, the more powerful the system becomes.

It should be further recognized that other applications besides the detection of spam e-mail include the detection of viruses, and the identification of copyrighted material
20 which are transmitted via the network. Moreover, it should be recognized that the algorithm for processing digital identifiers and the data store DS30 are not static, but can be adjusted to look for other characteristics of the message or data which is being tested besides frequency.

Hence, the system allows for leveraging between the number of first tier systems
25 or clients coupled to the database to provide a filtering system which utilizes a limited amount of bandwidth while still providing a confidential and powerful e-mail filter. It should be further recognized that the maintainer of the second and third tier systems may generate revenue for the service provided by charging a fee for the service of providing the second tier system process.

30 Still further, the system can collect and distribute anonymous statistical data about the content classified. For example, where e-mail filtering is the main application of the system, the system can identify the percentage of total e-mail filtered which constitutes spam, where such e-mail originates, and the like, and distribute it to interested parties for a fee or other compensation.

35

WO 01/46872

PCT/US00/42832

CLAIMS

What is claimed is:

1. A file content classification system comprising:
5 a digital ID generator;
an ID appearance database coupled to receive IDs from the ID generator; and
a characteristic comparison routine identifying the file as having a characteristic
based on ID appearance in the appearance database.
- 10 2. The content classification system of claim 1 wherein said ID generator comprises
a hashing algorithm.
3. The content classification system of claim 2 wherein said hashing algorithm is the
MD5 hashing algorithm.
- 15 4. The content classification system of claim 1 wherein said ID appearance database
tracks the frequency of appearance of a digital ID.
5. The content classification system of claim 1 further including a plurality of digital
20 ID generators on different systems all coupled to and providing IDs to said ID appearance
database.
6. The content classification system of claim 5 wherein said plurality of digital ID
generators are coupled to said database via a combination of public and private networks.
- 25 7. The content classification system of claim 6 wherein said database is coupled to
an intermediate server which is coupled to said plurality of generators.
8. The content classification system of claim 6 wherein said intermediate server is a
30 web server.
9. The content classification system of claim 1 wherein said characteristic comprises
junk e-mail and said characteristic is defined by a frequency of appearance of a digital ID.
- 35 10. A method for identifying a characteristic of a data file, comprising:

WO 01/45872

PCT/US00/42832

- generating a digital identifier for the data file and forwarding the identifier to a processing system;
- determining whether the forwarded identifier matches a characteristic of other identifiers; and
- 5 processing the email based on said step of determining.
11. The method of claim 10 wherein said step of generating comprises hashing at least a portion of the data file.
- 10 12. The method of claim 11 wherein said step of hashing comprises using the MD5 hash.
13. The method of claim 11 wherein said step of generating comprises hashing multiple portions of the data file.
- 15 14. The method of claim 10 wherein said data file is an email message and said step of determining comprises determining whether said email is spam.
- 15 15. The method of claim 10 wherein said step of determining identifies said e-mail as spam by tracking the rate per unit time a digital ID is generated.
- 20 16. The method of claim 10 wherein said step of generating comprises generating IDs at a plurality of source systems all coupled via a network to at least one processing system performing the determining step.
- 25 17. The method of claim 16 wherein said step of processing comprises instructing said plurality of source systems to perform an action with the email based on said determining step.
- 30 18. A method of filtering an email message, comprising:
processing the message to provide a digital identifier;
comparing the digital identifier to a characteristic database of digital identifiers to determine whether the message has said characteristic; and
processing the message based on said step of comparing.
- 35

WO 01/46872

PCT/US00/42832

19. The method of claim 18 wherein said step of processing occurs on at least one first system, and said step of comparing occurs on a second system.
20. The method of claim 19 wherein said step of processing occurs on a plurality of first systems.
21. The method of claim 19 wherein said at least one first system and second system are coupled by the Internet.
22. The method of claim 18 wherein said step of comparing comprises determining the frequency of a particular ID occurring in a time period, classifying said ID as having a characteristic, and comparing digital identifiers to said classified IDs.
23. A file content classification system, comprising:
 a first system having a file to be classified;
 an file ID generator on the first system;
 a database on a second system coupled to the ID generator to receive IDs generated by the ID generator;
 a comparison routine on the second system classifying the ID relative to the database as meeting or not meeting a characteristic.
24. The system of claim 23 including a plurality of first systems each including a respective file ID generator coupled to the database on the second system.
25. The system of claim 24 wherein the plurality of first systems is coupled to the second system via the Internet.
26. The system of claim 25 wherein the second system comprises a web server interface system and a database system, wherein the database system is isolated from the Internet by the web server system.
27. A content classification system for a first and second computer coupled by a network, comprising:
 a client agent file identifier generator on the first computer; and
 a server comparison agent and data-structure on the second computer receiving

WO 01/46872

PCT/US00/42832

identifiers from the client agent and providing replies to the client agent;
wherein the client agent processes the file based on replies from the server
comparison agent.

- 5 28. A method for providing a service on the Internet, comprising:
collecting data from a plurality of systems having a client agent on the Internet to
a server having a database;
characterizing the data received relative to information collected in the database;
and
10 transmitting a content identifier to the client agent.
29. The method of claim 28 wherein said step of collecting comprises collecting a
digital identifier for a data file.
- 15 30. The method of claim 28 wherein said data file is an e-mail.
31. The method of claim 29 wherein said step of characterizing comprises:
tracking the frequency of the collection of a particular identifier;
characterizing the data file based on said frequency;
20 storing the characterization; and
comparing collected identifiers to the known characterization.

WO 01/46872

1 / 2

PCT/US00/42832

Figure 1



Figure 2

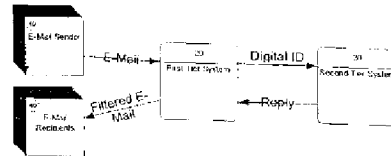
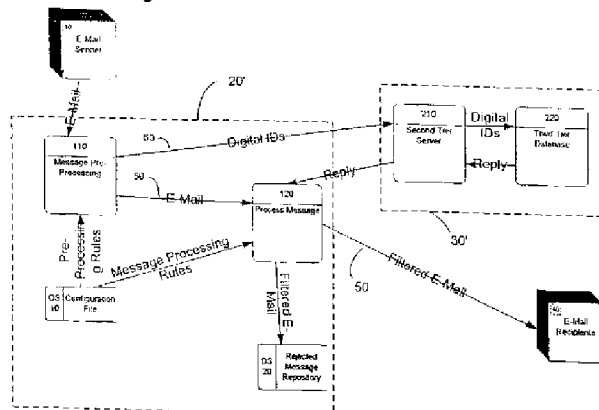


Figure 3



【国際調査報告】

INTERNATIONAL SEARCH REPORT		International application No. PCT/US00/42833
A. CLASSIFICATION OF SUBJECT MATTER IPC(Cl.) : G06F 17/30 US CL. : 707/9, 104; 709/201, 204 According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) I.S. : 707/6, 9, 104, 709/201, 202, 204, 225 Documentation searched other than minimum documentation in the extent that such documents are included in the fields searched Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) EAST		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,884,033 A (DUVAL, et al) 26 March 1999 (16.03.1999), ALL.	1-31
X, P	US 6,052,709 A (PAUL) 18 April 2000 (18.04.2000), All	1-31
<input type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/> See patent family search.		
* Symbolic references of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent published on or after the international filing date "L" document which may serve either as priority document or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "C" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is considered with one or more other such documents, such combinations being combinations previously skilled in the art "A" document of the same patent family		
Date of the actual completion of the international search 26 April 2001 (26.04.2001)		Date of mailing of the international search report 14 MAY 2001
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box 977 Washington, D.C. 20513 Facsimile No. (703) 205-5230		Authorized officer Uyen T Le <i>James R. Matthews</i> Telephone No. 963-9000

Form PCT/ISA/210 (second sheet) (July 1993)

フロントページの続き

(81)指定国 AP(GH,GM,KE,LS,MW,MZ,SD,SL,SZ,TZ,UG,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AT,BE,CH,CY,DE,DK,ES,FI,FR,GB,GR,IE,IT,LU,MC,NL,PT,SE,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AT,AU,AZ,BA,BB,BG,BR,BY,BZ,CA,CH,CN,CR,CU,CZ,DE,DK,DM,DZ,EE,ES,FI,GB,GD,GE,GH,GM,HU,ID,IL,IN,IS,JP,KE,KG,KP,KR,KZ,LC,LK,LR,LS,LT,LU,LV,MA,MD,MG,MK,MN,MW,MX,MZ,NO,NZ,PL,PT,RO,RU,SD,SE,SG,SI,SK,SL,TJ,TM,TR,TT,TZ,UA,UG,UZ,VN,YU,ZA,ZW

(74)代理人 100082005

弁理士 熊倉 禎男

(74)代理人 100065189

弁理士 穴戸 嘉一

(74)代理人 100096194

弁理士 竹内 英人

(74)代理人 100074228

弁理士 今城 俊夫

(74)代理人 100084009

弁理士 小川 信夫

(74)代理人 100082821

弁理士 村社 厚夫

(74)代理人 100086771

弁理士 西島 孝喜

(74)代理人 100084663

弁理士 箱田 篤

(72)発明者 ペイス マーク レイモンド

アメリカ合衆国 カリフォルニア州 94402 サン マテオ フィフティーンズ アベニュー
42

(72)発明者 タリー ブルックス キャッシュ

アメリカ合衆国 カリフォルニア州 94402 サン マテオ フィフティーンズ アベニュー
42

F ターム(参考) 5K030 GA15 HA06 LD20